Copy-Number-Variation and Copy-Number-Alteration Region Detection by Cumulative Plots

Wentian Li*, Annette Lee, Peter K Gregersen

The Robert S. Boas Center for Genomics and Human Genetics, The Feinstein Institute for Medical Research for Medical Research, North Shore LIJ Health System, Manhasset, 350 Community Drive, NY 11030, USA.

preprint version of *BMC Bioinformatics*, 10(suppl 1):S67 (2009)

Abstract

Background: Regions with copy number variations (in germline cells) or copy number alteration (in somatic cells) are of great interest for human disease gene mapping and cancer studies. They represent a new type of mutation and are larger-scaled than the single nucleotide polymorphisms. Using genotyping microarray for copy number variation detection has become standard, and there is a need for improving analysis methods. **Results:** We apply the cumulative plot to the detection of regions with copy number variation/alteration, on samples taken from a chronic lymphocytic leukemia patient. Two sets of whole-genome genotyping of 317k single nucleotide polymorphisms, one from the normal cell and another from the cancer cell, are analyzed. We demonstrate the utility of cumulative plot in detecting a 9Mb (9×10^6 bases) hemizygous deletion and 1Mb homozygous deletion on chromosome 13. We also show the possibility to detect smaller copy number variation/alteration regions below the 100kb range. **Conclusions:** As a graphic tool, the cumulative plot is an intuitive and a scale-free (window-less) way for detecting copy number variation/alteration regions, especially when such regions are small.

*Corresponding author: wli@nslij-genetics.org

Background

Most efforts in genetic mapping of human diseases focus on single-nucleotide-polymorphism (SNP): individual nucleotide base that may differ from one person to another. If the cause of a polymorphism is due to diverging paths in population genetic history, such as in multiple ethnic groups, it can be used as an ancestry or ethnic identity marker[1]. If the polymorphism is a functional mutation (non-synonymous or promoter-region polymorphism)[2] underlying a human disease, then it is the focus of attention in case-control genetic analyses[3].

A new type of genetic polymorphism emerged recently as another source of mutation that may lead to human diseases: the copy number variation (CNV) (for literature on CNV, see an online bibliography [4]). Local duplication and deletion events occuring at kb (10^3 bases) or Mb (10^6 bases) scales are the cause of CNV. If these events occurred in prior generations, CNV can be treated as a genetic marker whose transmission might be traced in studying the disease-status correlation. These events can also occur in the current generation, as *de novo* mutations.

Similar duplication and deletion events also occur in somatic cells, leading to copy number alteration (CNA). Besides the link between CNA and cancers studied before[5], an early CNV-disease association was reported on Charcot-Marie-Tooth disease[6], in inherited neurological disorder. In the past year or two, the number of reports on association of CNV with human diseases increased dramatically, especially for psychiatric disorders such as Schizophrenia[7, 8, 9], bipolar[10], and for brain developmental disorder such as Autism[11, 12, 13, 14, 15]. These diseases have long been evading genetic dissection, and the CNV link offers new optimism for our ultimate understanding of these diseases.

The technology for CNV detection evolves from Mb-level comparative genomic hybridization (CGH) to higher-resolution array-based CGH[16]. Genotyping array whose original goal is to genotype individual SNPs, has increasingly been used for CNV detection[17, 18, 19, 20]. There are two relevant pieces of information from a genotyping array data for the purpose of CNV detection.

The first is the ratio of intensity reading of alleles for a sample to that from a reference group of normal samples. If the ratio is larger than 1, there are more copies of piece of

DNA in the sample than normal (which is 2 copies). If the ratio is less than 1, it indicates a deletion. The second signal is the genotype. Deletion of one of the chromosomes leads to a run of homozygosity for all SNPs in the region, though run of homozygosity can also be due to inbreeding[21, 22]. The homozygosity property of one-copy deletion is well exploited in detecting loss-of-heterozygosity in CNA of cancer cells[23].

CNV detection using genotype microarray data relies on these two sequences: if the intensity ratio deviates from the normal value of 1 for a chromosome region with a consistent value, it can be a CNV region. Similarly, if a run of homozygosity is observed in a region, it could indirectly indicate a copy-number deletion. A CNV region detection is more convincing if CNV signals exhibited by both sequences overlap in a common region.

Methods for calling CNV regions can be roughly classified into two types. The first type is straightforward: a CNV detection is claimed when the log-ratio value is significantly deviated from 0[24]. The problem with this method is that the threshold for calling CNV varies greatly from platform to platform, from study to study, and a comparative investigation is urgently needed[16]. The second type uses hidden Markov models (HMM), where the underlying CNV status is the hidden variable, and the log-ratio and genotype sequences are the two observed variables[25, 26, 27, 28, 29, 30]. One advantage of the HMM framework is that it can incorporate information from both sequences at once.

When the parameter settings in a HMM are fixed, HMM is a stationary (homogeneous) process along a chromosome. There is one parameter in HMM which controls the transition probability from the (hidden) CNV state to non-CNV state. That parameter can also be transformed to the characteristic size for CNV region[28]. What if the CNV regions do not have a characteristic size, or equivalently, the length distribution is not exponential? In that case, CNV-calling methods that do not require stationarity are preferred.

The guanine-cytosine content (GC%) in DNA sequences has been a focus of non-stationary, non-Markov, long-range-correlated modeling for more than twenty years[31, 32, 33]. It is well acknowledged that the hierarchical pattern of GC%-domains within GC%-domains is possible[34, 35]. In order to detect both small and large GC-homogeneous domains, one applies methods that do not preset a characteristic scale. One such method is the recursive segmentation that adopts a divide-and-conquer approach[36]. Another is the cumulative plot.

Cumulative plot is a graphic display of sequence information such that trend in a region becomes more visible and obvious. It is a window-less method because no characteristic scale needs to be specified, although a window can be imposed to a plot when all patterns within certain length scale are to be ignored. In DNA sequence context, such cumulative plots were called "DNA walk"[37, 38] or "Z curve"[39, 40]. The cumulative plot has also been widely used for detection of replication origin[41, 42]. To our knowledge, cumulative plots have not been applied to CNV/CNA detection. The purpose of this paper is not to provide a comprehensive comparison of various CNV/CNA-calling methods, but limited to the presentation and illustration of this new approach.

Results and discussion

Since our method applies equally to CNV and CNA data, here we examine the CNA pattern in a cancer patient with chronic lymphocytic leukemia (CLL)[43]. DNA samples from the patient's normal cell and that from the cancer cell are obtained and genotyped with 317,000 SNPs genomewide. Figure 1 shows the log-ratio and θ sequences (see Methods) for chromosome 13, where a 9Mb CNA region (deletion) in the cancer cell is clearly visible. A deletion region is characterized by a drop in log-ratio value, and an absence of heterozygosity. Our goal is to capture the same information using cumulative plots.

The left panel of Figure 2 shows the two cumulative plots corresponding to log-ratio sequence and homozygosity indicator sequence h, respectively. In the simplest version, at each new SNP, the curve moves up or down by an amount equal to the log-ratio value of that SNP, or by the presence of a homozygote $(+1)$ and a heterozygote (-1).

For a deletion region, the log-ratio value is consistently negative, and the first cumulative plot shows a drop; and genotype is consistently homozygous (also called run of homozygosity (ROH)), and the second cumulative plot shows a jump. However, from Figure 2 (left), even outside the CNA region, the first (second) cumulative plot continues to go down (up), reflecting a global abundancy of negative log-ratio over positive one (homozygotes over heterozygotes).

To remove the global or chromosome-wide average, we redraw a detrended cumulative plot (right panel of Figure 2) where the linear trend from the normal cell is subtracted from the

two cumulative plots. If the difference of global trends between the cancer and normal cell is an artifact, e.g., the poor DNA quality in cancer cell that leads to higher missing rate for genotype calling, thus seemingly lower heterozygote frequency, then the normal and cancer cumulative plots should be detrended separately. Without such an evidence, we use the linear trend in normal cell to detrend both samples to highlight the difference between the two.

Cumulative plots can be customized to pick any specially defined signal. Suppose we are mainly interested in regions with copy number equal to 1, i.e., hemizygous deletion. Such deletion region should exhibit two features: (1) log-ratio is equal to $\log(1/2) = -0.693147$ (as versus $\log(2/2) = 0$ in the normal situation); (2) homozygosity indicator equal to 1 (as versus to a mixture of -1's and 1's). For a SNP, we then define a "one deletion" indicator variable whose value is 1 if $-2 <$ log-ratio < -0.34657 (mid-point between -0.693147 and 0) and if its genotype is a homozygote, and the value -1 otherwise.

Figure 3 shows the cumulative plot for "one deletion" indicator variable, without or with detrending (by the linear trend in the normal sample). In both versions, the hemizygous deletion region can be seen clearly. Not only the cumulative plot detects the CNA region easily, but also it delineates the border of the deletion region accurately.

When deletion occurs in both chromosomes, called homozygous deletion, the copy number is equal to zero. For homozygous deletions, both A- and B-channel intensity (see Methods) is close to zero, and the $\log(r)$ is a large negative value. Because in the A- and B-channel plane (see Methods), these SNPs are near the origin, the angle θ can not be determined unambiguously. This leads to a broad distribution of θ values between 0 and 1, as can be seen from Figure 1 (top).

We define a "two deletions" indicator variable whose value is 1 if the log-ratio is < -2; and the value is -1 otherwise. Note that the genotype information is not used. Figure 4 shows the cumulative plot for the "two deletions" indicator variable for chromosome 13. One homozygous deletion region with $\sim$ 1Mb is clearly identified immediately adjacent to the 9Mb hemizygous deletion region.

The 9Mb deletion on chromosome 13 in our CLL sample, which was one of the known common deletions for this disease[44], represents an example of easy detection of CNA/CNV region, because the difference between the normal and cancer cell for both log-ratio and geno-

type sequence is already obvious from the raw data (Figure 1). The advantage of cumulative plot is perhaps its ability to detect CNA/CNV region of smaller sizes.

Figure 5 shows the example of chromosome 6 of our sample where there is no large-scaled CNA region. The log-ratio and genotype sequence look almost identical between the normal and the cancer cell. The cumulative plot for the "one deletion" indicator variable shows that there are +400 more SNPs in the cancer cell than in the normal cell to have the one deletion signal (the "two deletion" cumulative plot is not shown because the signal is mostly absent along the chromosome). However, these SNPs are distributed throughout the chromosome, instead of forming clusters, and we still do not have strong evidence that the cancer sample has more micro deletion regions as compared to the normal sample.

In order to explore the possible existence of smaller CNA regions, we pick the longest ROH region (roughly 4Mb) and view it with cumulative plots. Figure 6 (left) shows the un-detrended cumulative plot for the one-deletion indicator variable in this region. A clear hemizygous deletion region should show up as a jump in the cumulative plot. However, the tendency within this ROH region is downward instead of upward. In other words, although all genotypes in this region are homozygous, the log-ratio mostly fails the < -0.34657 criterion.

The failure in detecting hemizygous deletion at the Mb scale does not necessarily prevent its possible existence at a smaller length scale. The right panel of Figure 6 shows a 200kb sub-region (marked in Figure 6 (left)) that contains a 36kb region with an upward trend in the cumulative plot. A zooming into any small region in a cumulative plot enables it to detect CNA/CNV regions with ever smaller sizes.

It was previously suggested that run of homozygosity can be a sequence feature that is associated with certain human diseases[45]. We see here that ROH is only a partial indicator for a CNA/CNV region. The longest ROH on chromosome 6 in our sample only shows some weak evidence in a much narrower region for one-deletion CNA. Considering both ROH and log-ratio sequence is clearly better than considering ROH alone. Although ROH may still be biologically meaningful, as it could reflect a copy-neutral loss-of-heterozygosity event, one has to obtain extra evidence to exclude population genetics events such as inbreeding as the true cause.

The pairing of the normal and the cancer sample is not essential to our method. In Figures

3,4,6, the CNA regions can be identified by cancer sample (the blue curve) alone. However, the comparison with the normal sample provides supporting evidence that deletion only occurs in the cancer cell and not in the normal cell.

When SNPs along a chromosome are not evenly distributed, it may not be appropriate to move one step per SNP in the cumulative plot. For example, if multiple SNPs are in strong linkage disequilibrium in a densely typed region, the indicator variable values are positively correlated, and a sequence of +1 values is partially a consequence of their correlation, not as a series of independent evidences for CNA/CNV. We can adjust for this correlation by calculating the probability ratio α (see Method) in favor for concordant genotypes between neighboring SNPs, as compared to the average. If $\alpha > 1$, we discount a +1 (or -1) movement by dividing the α value. For the chromosome 13 data, α is in a very narrow range of (0.9921, 1.0002). Because the probability ratio in favor of concordant homozygotes is so close to 1, the adjusted cumulative plot is indistinguishable from the original cumulative plot.

So far the delineation of an upward trend in the cumulative plot is determined by visual inspection. Segmentation programs can be developed to carry out the delineation automatically. In particular, one may move along the cumulative plot, calculate the slope from the start point to the moving position, then from the moving position to the end point. The position that maximizes the difference of the two slopes is chosen, leading to the first segmentation. This segmentation can be carried out recursively similar to the method described in [36].

Finally, for case-control analysis using CNV, one deals with two groups of samples[46]. In this situation, cumulative plot can be first applied to each individual person to identify the CNV/CNA region. Then, chromosomes can be partitioned into equal-sized windows and the frequency of CNV/CNA-containing window in the case group is compared to that in the control group for a statistical test.

Conclusions

We have shown here that cumulative plots of an indicator variable derived from the log-ratio and SNP genotype sequence can easily identify CNV or CNA regions. We illustrate the procedure for hemizygous deletion (copy number equal to 1) and homozygous deletion

(copy number equal to 0) using samples taken from a chronic lymphocytic leukemia patient. Although CNV/CNA regions at the Mb scale can also be detected by viewing the raw data, cumulative plot is able to delineate the borders with higher degree of accuracy. Another advantage of cumulative plot is perhaps in detecting smaller CNV/CNA regions, such as those in the range of 10kb-100kb, as it is a scale-free approach that does not require a fixing of the window size. Cumulative plot is simple enough that no special-purpose program is needed for its use except a graphic routine: for example, all results shown here are obtained by the general statistical package R[47].

Methods

log-ratio and genotype data

In a two-channel (two-color) SNP genotyping microarray, the A- and B- channel (A- and B-allele) intensity reading is recorded. These two intensities are normalized by reference intensity values which are obtained by averaging many normal samples. Each SNP can be represented by a point in the (x, y) plane where x, y are the normalized A- and B-channel intensity. The polar coordinate of the point is $r = \sqrt{x^2 + y^2}$ and $\theta = tan^{-1}(y/x)$[48]. Log(r) is the "log ratio" value that provides a copy-number information, and θ provides a genotype information, where $\theta = 0, 1$ correspond to two homozygotes, and $\theta = 0.5$ corresponds to the heterozygote. Note: (1) r value depends on a group-averaged reference level, and this information is provided by the array-maker company. (2) Although r and θ is in principle independent, there could be weak correlation between them. Our starting point are the two sequences of log(r) and discretized θ values (i.e. genotype) along a chromosome.

Cumulative plots for log-ratio and homozygosity sequence

The r and θ variable is transformed by: log-ratio = $log(r)$ and homozygosity indicator $h = 4 \times |\theta - 0.5| - 1$. For heterozygotes, h is close to -1, and for two homozygotes, h is close to 1. Denote the i-th SNP's log-ratio and homozygosity indicator as $\log(r_i)$ and h_i. The (original)

cumulative plots of these two sequences are:

$$\begin{aligned} cumu.log.ratio_j &= \sum_{i=1}^{j} \log(r_i) \\ cumu.h_j &= \sum_{i=1}^{j} h_i \end{aligned} \quad (1)$$

A cumulative plot can be detrended such that the first and the last SNP are on the same horizontal line. The purpose of this detrending is to remove the chromosome-wide bias so that regional deviations are highlighted. In our normal and cancer cell from the same individual example, we detrend the normal sample by subtracting the linear function $a + bx_i$, where x_i is the Mb position of the ith SNP, N is the number of SNPs, and

$$\begin{aligned} b &= \frac{cumu.log.ratio_N - cumu.log.ratio_1}{x_N - x_1} \\ a &= cumu.log.ratio_N - bx_N. \end{aligned} \quad (2)$$

To highlight the difference between the cancer cell and the normal cell, we use the a and b obtained from the normal cell to detrend the cumulative plot for the cancer cell.

Cumulative plots corrected by spacing between neighboring SNPs

When SNPs are not distributed evenly along a chromosome, one may consider correcting the effect of inhomogeneous correlation between neighboring SNPs. We first calculate the probability of a neighboring SNP of a homozygous SNP to be also homozygous due to the correlation between them. This calculation is carried out by the Haldane's map[49].

Haldane's map relates the number of recombinations within a chromosomal interval M and the probability of observing a recombinant between the two end points R:

$$R = \frac{1 - exp(-2M)}{2}. \quad (3)$$

The unit of M is Morgan, which is roughly equal to 100Mb (or 1 centi Morgan is equal to 1 Mb[50]). The probability of observing a non-recombinant is $1 - R$.

Denote p_{same} the probability that one homozygous SNP is followed by another homozygous SNP that is M genetic distance apart. Since Haldane formula is applicable to haplotype, or a single copy of a chromosome, for two copies of a chromosome, $p_{same} = (1-R)^2 \approx 1-2R = e^{-2M}$.

Suppose the average spacing between two neighboring SNPs is $\overline{M}$. For a neighboring SNP pair whose spacing $M < \overline{M}$, it is more likely for both SNPs to be homozygous than the average, by a probability ratio of $\alpha = p_{same}/\overline{p}_{same} = e^{-2(M-\overline{M})}$, and the cumulative plot for the homozygosity indicator variable can be adjusted by dividing that ratio:

$$cumu.h_j = \sum_{i=1}^{j} h_i/\alpha_i = \sum_{i=1}^{j} h_i e^{2(M_{i-1,i}-\overline{M})}. \quad (4)$$

We assume that p_{same} is calculated in the same way as for other indicator variables, meaning CNV/CNA of a particular type is maintained at the neighboring SNP by the same probability e^{-2M}, and the above formula can be used to correct other cumulative plots. Note that transition probability from one genotype in a SNP to another genotype in the neighboring SNP can also be estimated from the HapMap data.

Authors contributions

W.L. designed the method, carried out the analysis, and wrote the manuscript; A.L. genotyped the samples; P.K.G. proposed the CNV study of chronic lymphocytic leukemia.

List of abbreviations

CGH: comparative genomic hybridization **CLL:** chronic lymphocytic leukemia **CNA:** copy number alterations **CNV:** copy number variations **GC%:** guanine and cytosine contents **HMM:** hidden Markov models **ROH:** run of homozygosity **SNP:** single nucleotide polymorphism

Acknowledgements

We thank Nick Chiorazzi for providing the CLL sample, and Pedro Bernaola-Galván, José Oliver for discussions on segmentation methods.

References

[1] Rosenberg N, Li L, Ward R, Pritchard J: **Informativeness of genetic markers for inference of ancestry**. *Am. J. Hum. Genet.* 2003, **73**:1402–1422.

[2] Botstein D, Risch N: **Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease**. *Nature Genet.* 2003, **33**:228–237.

[3] Li W: **Three lectures on case-control genetic association analysis**. *Brief. Bioinform.* 2008, **9**:1–13.

[4] **An online bibliography on copy number variations** [*http://www.nslij-genetics.org/cnv/*].

[5] Kallioniemi A, Kallioniemi O, Sudar D, Rutovitz D, Gray J, Waldman F, Pinkel D: **Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors**. *Science* 1992, **258**:818–821.

[6] Lupski J, de Oca-Luna R, Slaugenhaupt S, Pentao L, Guzzetta V, Trask B, Saucedo-Cardenas O, Barker D, Killian J, Garcia C, Chakravarti A, Patel P: **DNA duplication associated with Charcot-Marie-Tooth disease type 1A**. *Cell* 1991, **66**:219–232.

[7] Sutrala S, Goossens D, Williams N, Heyrman L, Adolfsson R, Norton N, Buckland P, Del-Favero J: **Gene copy number variation in schizophrenia**. *Schiz. Res.* 2003, **96**:1–3.

[8] Walsh T, McClellan JM, McCarthy SE, Addington AM, Pierce SB, Cooper GM, Nord AS, Kusenda M, Malhotra D, Bhandari A, Stray SM, Rippey CF, Roccanova P, Makarov V, Lakshmi B, Findling RL, Sikich L, Stromberg T, Merriman B, Gogtay N, Butler P, Eckstrand K, Noory L, Gochman P, Long R, Chen Z, Davis S, Baker C, Eichler EE, Meltzer PS, Nelson SF, Singleton AB, Lee MK, Rapoport JL, King MC, Sebat J: **Rare structural variants disrupt multiple genes in neurodevelopmental pathways in Schizophrenia**. *Science* 2008, **358**:667–675.

[9] Xu B, Roos J, Levy S, van Rensburg E, Gogos J, Karayiorgou M: **Strong association of de novo copy number mutations with sporadic schizophrenia**. *Nature Genet.* 2008, **40**:880–885.

[10] Wilson G, Flibotte S, Chopra V, Melnyk B, Honer WG, Holt R: **DNA copy-number analysis in bipolar disorder and schizophrenia reveals aberrations in genes involved in glutamate signaling**. *Hum. Mol. Genet.* 2006, **15**:743–749.

[11] Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee Y, Hicks J, Spence S, Lee A, Puura K, Lehtimki T, Ledbetter D, Gregersen P, Bregman J, Sutcliffe J, Jobanputra V, Chung W, Warburton D, King M, Skuse D, Geschwind D, Gilliam T, Ye K, Wigler M: **Strong association of de novo copy number mutations with autism**. *Science* 2007, **316**:445–449.

[12] Szatmari P, Paterson AD, Zwaigenbaum L, Roberts W, Brian J, Liu XQ, Vincent JB, Skaug JL, Thompson AP, Senman L, Feuk L, Qian C, Bryson SE, Jones MB, Marshall CR, Scherer SW, Vieland VJ, Bartlett C, Mangin LV, Goedken R, Segre A, Pericak-Vance MA, Cuccaro ML, Gilbert JR, Wright HH, Abramson RK, Betancur C, Bourgeron T, Gillberg C, Leboyer M, Buxbaum JD, Davis KL, Hollander E, Silverman JM, Hallmayer J, Lotspeich L, Sutcliffe JS, Haines JL, Folstein SE, Piven J, Wassink TH, Sheffield V, Geschwind DH, Bucan M, Brown WT, Cantor RM, Constantino JN, Gilliam TC, Herbert M, Lajonchere C, Ledbetter DH, Lese-Martin C, Miller J, Nelson S, Samango-Sprouse CA, Spence S, State M, Tanzi RE, Coon H, Dawson G, Devlin B, Estes A, Flodman P, Klei L, McMahon WM, Minshew N, Munson J, Korvatska E, Rodier PM, Schellenberg GD, Smith M, Spence MA, Stodgell C, Tepper PG, Wijsman EM, Yu CE, Rog B, Mantoulan C, Wittemeyer K, Poustka A, Felder B, Klauck SM, Schuster C, Poustka F, Bölte S, Feineis-Matthews S, Herbrecht E, Schmtzer G, Tsiantis J, Papanikolaou K, Maestrini E, Bacchelli E, Blasi F, Carone S, Toma C, Engeland HV, Jonge MD, Kemner C, Koop F, Langemeijer M, Hijmans C, Staal WG, Baird G, Bolton PF, Rutter ML, Weisblatt E, Green J, Aldred C, Wilkinson JA, Pickles A, Couteur AL, Berney T, McConachie H, Bailey AJ, Francis K, Honeyman G, Hutchinson A, Parr JR, Wallace S, Monaco AP,

Barnby G, Kobayashi K, Lamb JA, Sousa I, Sykes N, Cook EH, Guter SJ, Leventhal BL, Salt J, Lord C, Corsello C, Hus V, Weeks DE, Volkmar F, Tauber M, Fombonne E, Shih A, Meyer KJ: **Mapping autism risk loci using genetic linkage and chromosomal rearrangements**. *Nature Genet.* 2007, **39**:319–328.

[13] Weiss LA, Shen Y, Korn JM, Arking DE, Miller DT, Fossdal R, Saemundsen E, Stefansson H, Ferreira MA, Green T, Platt OS, Ruderfer DM, Walsh CA, Altshuler D, Chakravarti A, Tanzi RE, Stefansson K, Santangelo SL, Gusella JF, Sklar P, Wu BL, Daly MJ: **Association between microdeletion and microduplication at 16p11.2 and Autism**. *New Eng. J. Med.* 2008, **358**:667–675.

[14] Marshall CR, Noor A, Vincent JB, Lionel AC, Feuk L, Skaug J, Shago M, Moessner R, Pinto D, Ren Y, Thiruvahindrapduram B, Fiebig A, Schreiber S, Friedman J, Ketelaars CE, Vos YJ, Ficicioglu C, Kirkpatrick S, Nicolson R, Sloman L, Summers A, Gibbons CA, Teebi A, Chitayat D, Weksberg R, Thompson A, Vardy C, Crosbie V, Luscombe S, Baatjes R, Zwaigenbaum L, Roberts W, Fernandez B, Szatmari P, Scherer SW: **Structural variation of chromosomes in autism spectrum disorder**. *Am. J. Hum. Genet.* 2008, **82**:477–488.

[15] Eichler E, Zimmerman AW: **A hot spot of genetic instability in Autism**. *New Eng. J. Med.* 2008, **358**:737–739.

[16] Carter N: **Methods and strategies for analyzing copy number variation using DNA microarrays**. *Nature Genet.* 2007, **39(suppl.)**:S16–S21.

[17] Huang J, Wei W, Zhang J, Liu G, Bignell GR, Stratton MR, Futreal PA, Wooster R, Jones KW, Shapero MH: **Whole genome DNA copy number changes identified by high density oligonucleotide arrays**. *Hum. Genomics* 2004, **1**:287–299.

[18] Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, Hangaishi A, Kurokawa M, Chiba S, Bailey DK, Kennedy GC, Ogawa S: **A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays**. *Cancer Res.* 2005, **65**:6071–6079.

[19] Newman TL, Rieder MJ, Morrison VA, Sharp AJ, Smith JD, Sprague LJ, Kaul R, Carlson CS, Olson MV, Nickerson DA, Eichler EE: **High-throughput genotyping of intermediate-size structural variation**. *Hum. Mol. Genet.* 2006, **15**:1159–1167.

[20] Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, Cheung SW, Shen RM, Barker DL, Gunderson KL: **High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping**. *Genome Res.* 2006, **16**:1136–1148.

[21] Gibson J, Morton N, Collins A: **Extended tracts of homozygosity in outbred human populations**. *Hum. Mol. Genet.* 2006, **15**:789–795.

[22] Simon-Sanchez J, Scholz S, Fung HC, Matarin M, Hernandez D, Gibbs JR, Britton A, Vrieze FD, Peckham E, Gwinn-Hardy K, Crawley A, Keen JC, Nash J, Borgaonkar D, Hardy J, Singleton A: **Genome-wide SNP assay reveals structural genomic variation, extended homozygosity and cell-line alterations in normal individuals**. *Hum. Mol. Genet.* 2007, **16**:1–14.

[23] Lin M, Wei L, Sellers W, Lieberfarb M, Wong W, Li C: **dChipSNP: significance curve and clustering of SNP-array-based loss-of-heterozygosity data**. *Bioinform.* 2004, **20**:1233–1240.

[24] Vermeesch JR, Melotte C, Froyen G, Vooren SV, Dutta B, Maas N, Vermeulen S, Menten B, Speleman F, Moor BD, Hummelen PV, Marynen P, Fryns JP, Devriendt K: **Molecular karyotyping: array CGH quality criteria for constitutional genetic diagnosis**. *J. Histochem. & Cytochem.* 2005, **53**:413–422.

[25] Zhao X, Li C, Paez JG, Chin K, Jänne P, Chen T, Girard L, Minna J, Christiani D, Leo C, Gray JW, Sellers WR, Meyerson M: **An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays**. *Cancer Res.* 2004, **64**:3060–3071.

[26] Fridlyand J, Snijders A, Pinkel D, Albertson D, Jain A: **Hidden Markov models approach to the analysis of array CGH data**. *J. Multivar. Anal.* 2004, **90**:132–153.

[27] Beroukhim R, Lin M, Park Y, Hao K, Zhao X, Garraway L, Fox E, Hochberg E, Mellinghoff I, Hofer M, Descazeaud A, Rubin M, Meyerson M, Wong W, Sellers W, Li C: **Inferring loss-of-heterozygosity from unpaired tumors using high-density oligonucleotide SNP arrays**. *PLoS Comp. Biol.* 2006, **2**:e41.

[28] Colella S, Yau C, Taylor J, Mirza G, Butler H, Clouston P, Bassett A, Seller A, Holmes C, Ragoussis J: **QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data**. *Nucl. Acids Res.* 2007, **35**:2013–2025.

[29] Wang K, Li M, Hadley D, Liu R, Glessner J, Grant S, Hakonarson H, Bucan M: **PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data**. *Genome Res.* 2007, **17**:1665–1674.

[30] Cahan P, Godfrey LE, Eis PS, Richmond TA, Selzer RR, Brent M, McLeod HL, Ley TJ, Graubert TA: **wuHMM: a robust algorithm to detect DNA copy number variation using long oligonucleotide microarray data**. *Nucl. Acids Res.* 2008, **36**:e41.

[31] Fickett J, Torney D, Wolf D: **Base compositional structure of genomes**. *Genomics* 1992, **13**:1056–1064.

[32] Li W, Kaneko K: **Long-range correlation and partial $1/f^{\alpha}$ spectrum in a noncoding DNA sequence**. *Europhys. Lett.* 1992, **17**:655–660.

[33] Melodelima C, Guéguen L, Piau D, Gautier C: **A computational prediction of isochores based on hidden Markov models**. *Gene* 2006, **385**:41–48.

[34] Li W: **Delineating relative homogeneous G+C domains in DNA sequences**. *Gene* 2001, **276**:57–72.

[35] Li W: **Are isochore sequences homogeneous?** *Gene* 2002, **300**:129–139.

[36] Bernaola-Galván P, Roman-Roldán R, Oliver J: **Compositional segmentation and long-range fractal correlations in DNA sequences**. *Phys. Rev. E* 1996, **53**:5181–5189.

[37] Berthelsen C, Glazier J, Skolnick M: **Global fractal dimension of human DNA sequences treated as pseudorandom walks**. *Phys. Rev. A* 1992, **45**:8902–8913.

[38] Peng CK, Buldyrev SV, Goldberger AL, Havlin S, Sciortino F, Simons M, Stanley HE: **Long-range correlations in nucleotide sequences**. *Nature* 1992, **356**:168–170.

[39] Zhang R, Zhang C: **Z Curves, an intuitive tool for visualizing and analyzing DNA sequences**. *J. Biomol. Struc. Dynamics* 1994, **11**:767–782.

[40] Zhang C, Zhang R: **An isochore map of the human genome based on the Z curve method**. *Gene* 2003, **317**:127–135.

[41] Grigoriev A: **Analyzing genomes with cumulative skew diagrams**. *Nucl. Acids Res.* 1998, **26**:2286–2290.

[42] Freeman J, Plasterer T, Smith T, Mohr S: **Patterns of genome organization in bacteria**. *Science* 1998, **279**:1827a.

[43] Chiorazzi N, Rai K, Ferrarini M: **Chronic lymphocytic leukemia**. *New Eng. J. Med.* 2005, **352**:804–815.

[44] Döhner H, Stilgenbauer S, Benner A, Leupolt E, Kröber A, Bullinger L, Döhner K, Bentz M, Lichter P: **Genomic aberrations and survival in chronic lymphocytic leukemia**. *New Eng. J. Med.* 2000, **343**:1910–1916.

[45] Lencz T, Lambert C, DeRosse P, Burdick K, Morgan T, Kane J, Kucherlapati R, Malhortra A: **Runs of homozygosity reveal highly penetrant recessive loci in schizophrenia**. *Proc. Natl. Acad. Sci.* 2007, **104**:19942–19947.

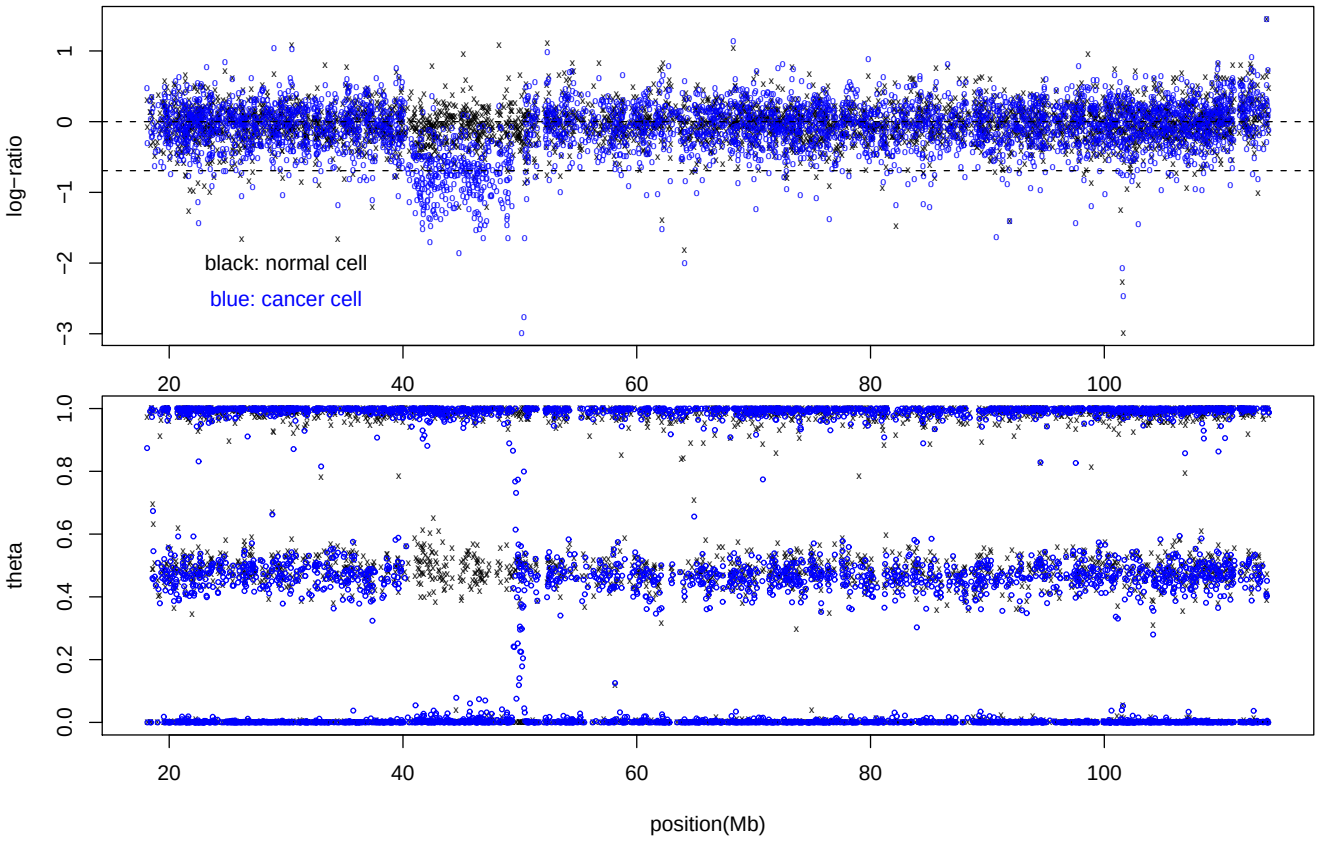
[46] Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, Clayton D, Hurles ME: **A robust statistical method for case-control association testing with copy number variation**. *Nature Genet.* 2008, **40**:1245–1252.

[47] **The R project for statistical computing** [*http://www.r-project.org/*].

[48] Peiffer D, Gunderson K: **SNP-CGH technologies for genomic profiling of LOH and copy number**. *Clinical Lab International* 2006, **May**.

[49] Ott J: *Analysis of Human Genetic Linkage*. Baltimore: The Johns Hopkins University Press, 3rd edition 1999.

[50] Ulgen A, Li W: **Comparing single-nucleotide-polymorphism marker-based and microsatellite marker-based linkage analyses**. *BMC Genet.* 2005, **6(suppl 1)**:S13.



Figures

Figure 1 - Log-ratio and genotype sequences for chromosome 13 in paired samples from a CLL patient

Log-ratio (top) and genotype θ (bottom) sequence for SNPs from chromosome 13 of two samples taken from the same cancer patient: black for normal cell and blue for cancer cell. For the log-ratio plot, the copy number of 2 level $\log(2/2) = 0$ and the copy number of 1 level $\log(1/2) = -0.693147$ are marked.

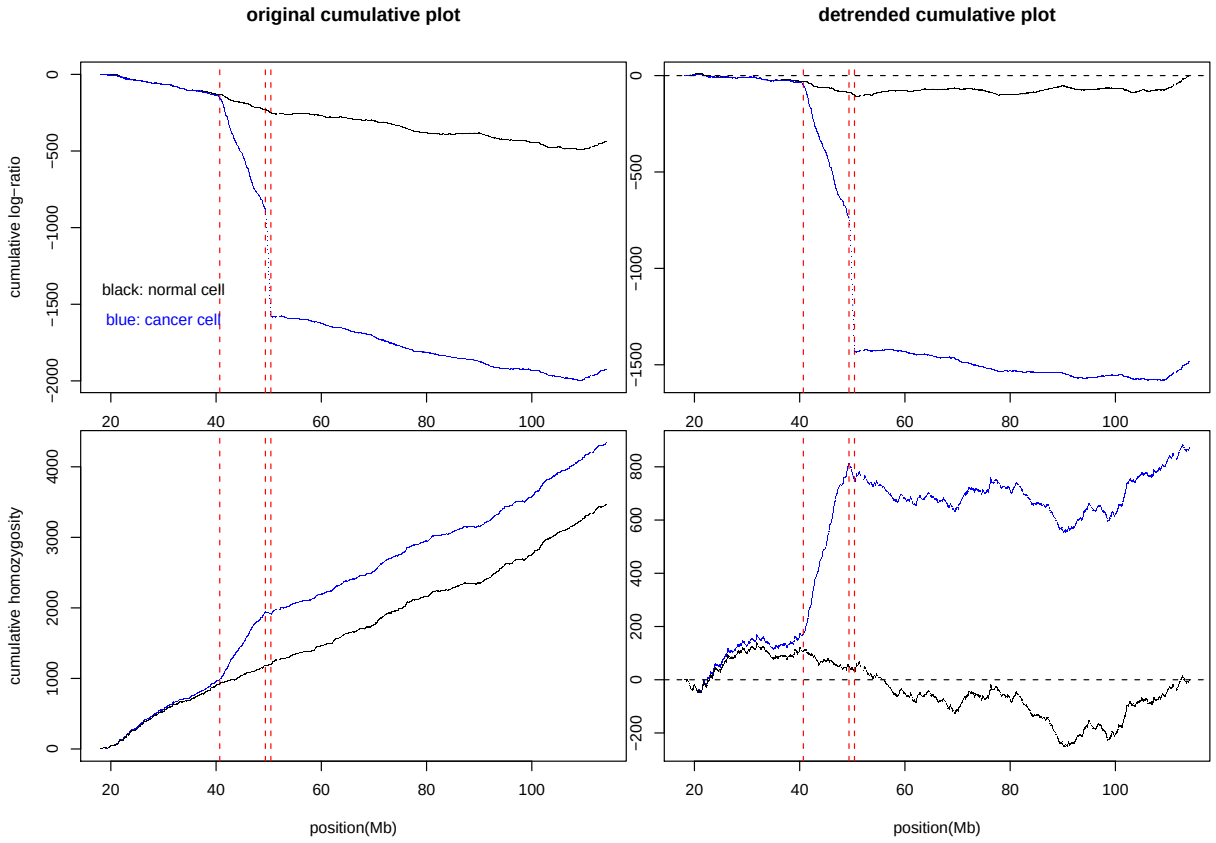


Figure 2 - Cumulative plot of log-ratio and homozygosity sequence

Cumulative plot and detrended cumulative plot for both the log-ratio sequence and the homozygosity indicator sequence (for the chromosome 13 data shown in Figure 1). Top: cumulative plots for log-ratio sequence. Bottom: cumulative plot for homozygosity sequence (1 for homozygote, -1 for heterozygote). Left: original cumulative plots. Right: detrended cumulative plots. The linear trend obtained from the normal sample is used to detrend both the normal and the cancer sample. Black for the normal cell and blue for the cancer cell. The 9MB hemizygous deletion and the neighboring 1Mb homozygous deletion region are marked by red lines.

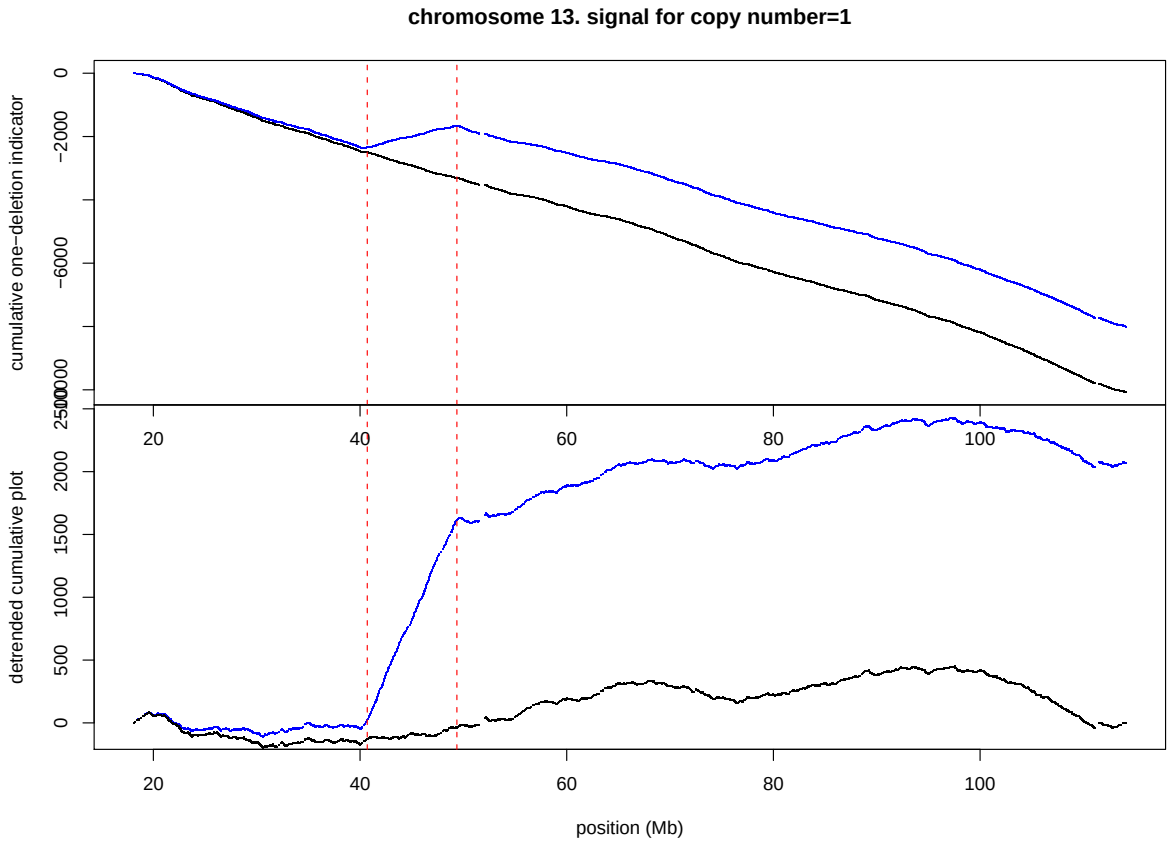


Figure 3 - Cumulative plot of the hemizygous deletion indicator variable

Cumulative plot (top) and detrended cumulative plot (bottom) for the 9Mb hemizygous deletion region on chromosome 13, using the "one deletion" indicator variable.

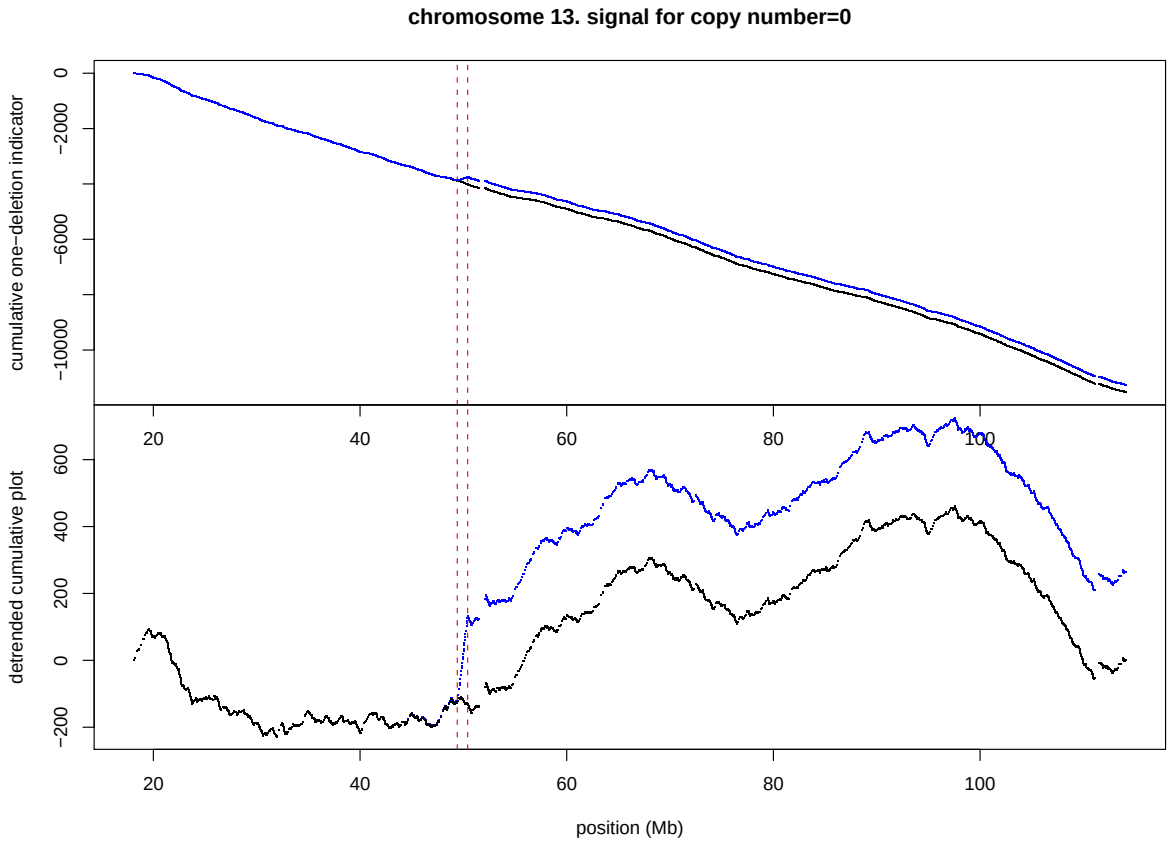


Figure 4 - Cumulative plot of the homozygous deletion indicator variable

Cumulative plot (top) and detrended cumulative plot (bottom) for the 1Mb homozygous deletion region on chromosome 13, using the "two deletions" indicator variable.

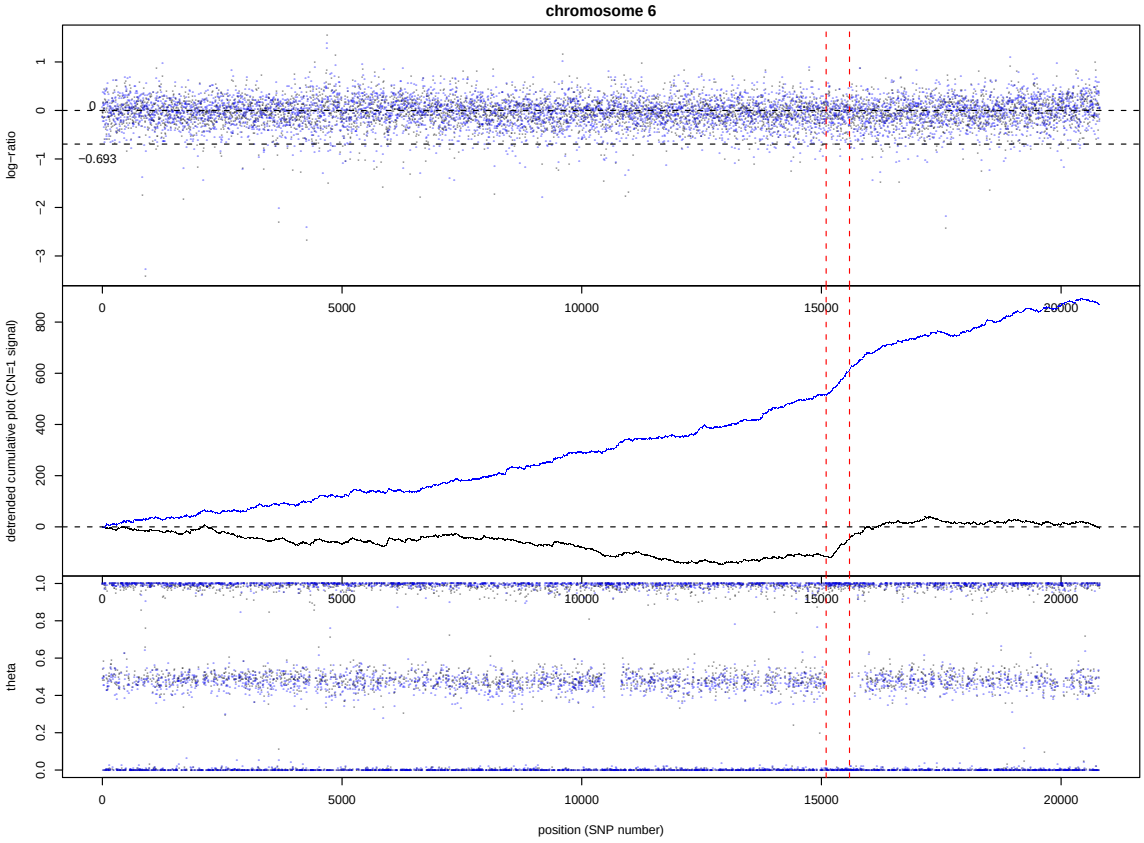


Figure 5 - Log-ratio and genotype sequences for chromosome 6 in paired samples from a CLL patient

The log-ratio sequence (top), genotype θ sequence (bottom), and the detrended cumulative plot for the "one deletion" indicator variable for SNPs on chromosome 6. Black and blue color refer to the normal and cancer cell sample taken from the same cancer patient. The largest run-of-homozygosity region is marked by red vertical lines. The copy number of 2 level $\log(2/2) = 0$ and the copy number of 1 level $\log(1/2) = -0.693147$ are marked in the log-ratio plot.

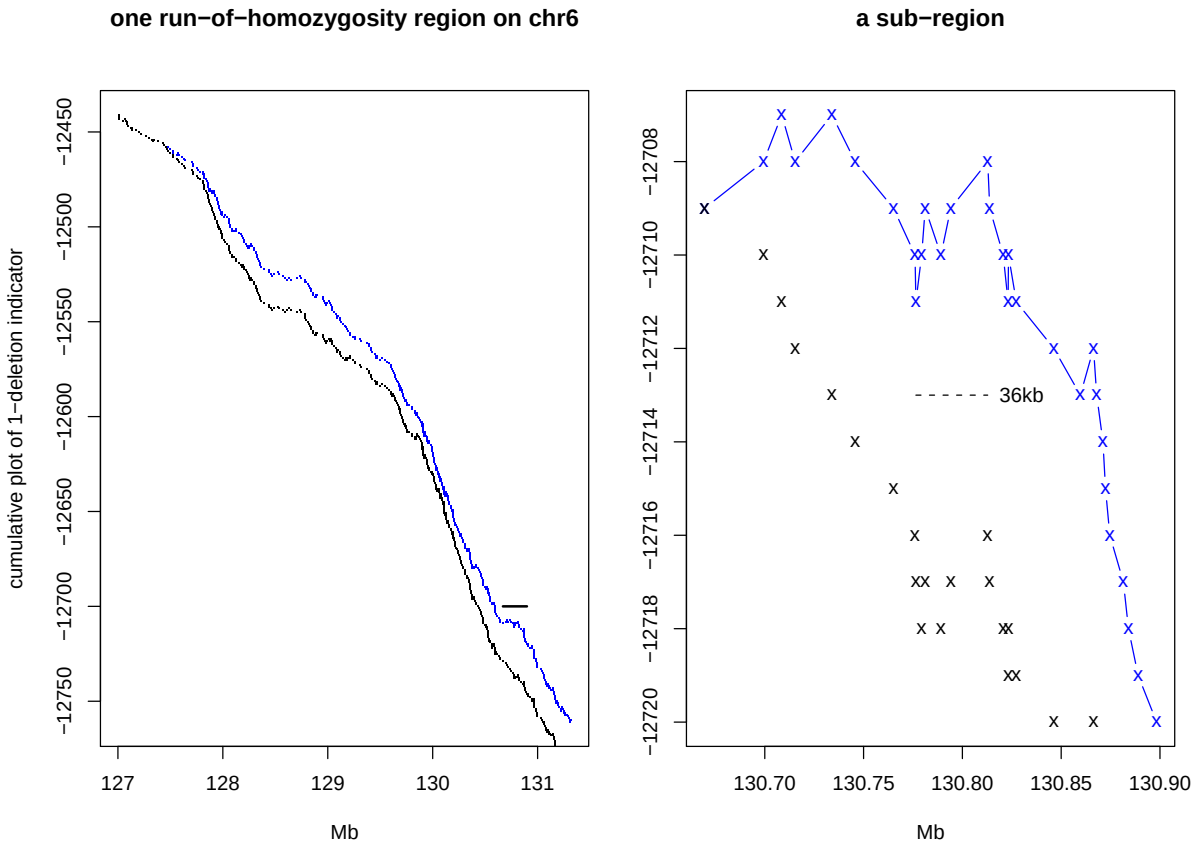


Figure 6 - Zoom in of smaller regions

Cumulative plots of "one deletion" indicator variable for the region marked in Figure 5 (left), and the sub-region marked by a horizontal bar on the left (right). Black and blue refer to the normal and the cancer sample.